Enhancement of esophageal speech using voice conversion techniques

Imen Ben Othmane, Joseph Di Martino, Kaïs Ouni

Enhancement of esophageal speech using voice conversion techniques

Imen Ben Othmane[1,2], *Joseph Di Martino*[2], *Kais Ouni*[1]

[1]Research Unit Signals and Mechatronic Systems, SMS, UR13ES49,
National Engineering School of Carthage, ENICarthage University of Carthage, Tunisia
[2]LORIA - Laboratoire Lorrain de Recherche en Informatique et ses Applications,
B.P. 239 54506 Vandœuvre-lès-Nancy, France

imen.benothmen@hotmail.fr, joseph.di-martino@loria.fr, kais.ouni@enicarthage.rnu.tn

Abstract

This paper presents a novel approach for enhancing esophageal speech using voice conversion techniques. Esophageal speech (ES) is an alternative voice that allows a patient with no vocal cords to produce sounds after total laryngectomy: this voice has a poor degree of intelligibility and a poor quality. To address this issue, we propose a speaking-aid system enhancing ES in order to clarify and make it more natural. Given the specificity of ES, in this study we propose to apply a new voice conversion technique taking into account the particularity of the pathological vocal apparatus. We trained deep neural networks (DNNs) and Gaussian mixture models (GMMs) to predict "laryngeal" vocal tract features from esophageal speech. The converted vectors are then used to estimate the excitation cepstral coefficients and phase by a search in the target training space previously encoded as a binary tree. The voice resynthesized sounds like a laryngeal voice i.e., is more natural than the original ES, with an effective reconstruction of the prosodic information while retaining, and this is the highlight of our study, the characteristics of the vocal tract inherent to the source speaker. The results of voice conversion evaluated using objective and subjective experiments, validate the proposed approach.

Index Terms: Esophageal speech, deep neural network, Gaussian mixture model, excitation, phase, KD-Tree.

1. Introduction

Speech is a spontaneous communication tool. Unfortunately, many people are unable to speak correctly. For example people who have undergone a total removal of their larynx (laryngectomees) due to laryngeal cancer or accident, cannot produce laryngeal speech sounds anymore. They need another process to produce speech sounds without a larynx. There process are for example the Artificial Larynx Transducer (ALT), the tracheo-esophageal (TE) prosthesis, or the esophageal speech.

Among them, ES is more natural when compared to the voice generated by the other process. However, the degradation of naturalness and the low intelligibility of esophageal speech is caused by some factors such as chaotic fundamental frequency, and specific noises.

Consequently, laryngeal speech is not comparable to esophageal speech.

In order to be able to integrate again laryngectomees to their individual, social and work activities, some different approaches have been proposed to enhance speech after laryngectomy surgery. In [1] and [2] a statical approach has been implemented for enhancing ES in order to increase its quality. Other approaches based on the transformation of the acoustic features, such as smoothing [3] or comb filtering [4] have been proposed. But it is difficult to improve ES by using those simple modification methods because the properties of acoustic features of esophageal speech are totally different from those of normal speech. In this paper, we propose to use voice conversion techniques to enhance ES. The goal of voice conversion (VC) is to transform the audio signal from a source speaker as if a target speaker had spoken it. We train DNNs and GMMs [20], [29], [5], [6] with the acoustic features of esophageal speech and those of normal speech. For realizing this training procedure two parallel corpora consisting of utterance-pairs of the source esophageal speech and the target normal speech are used.

This paper is structured as follows: section 2 presents the principle of the proposed technique; the obtained results from the realized tests and experiments are exhibited in section 3; section 4 presents conclusions with prospects.

2. Problem definition and related work

2.1. Esophageal speech

ES is characterized by low intelligibility, high noise perturbation and unstable fundamental frequency. When compared with laryngeal (normal) voice, ES is hoarse, has a low and chaotic F0 and therefore is difficult to understand.

Figure 1 shows an example of speech waveforms and spectrograms for normal and esophageal speech for the same sentence. We can observe that the acoustic features of normal speech are very different from those of esophageal speech.

Esophageal speech often includes some specific noisy perturbations produced through a process of generating excitation signals by releasing the air from the stomach and the esophagus afterwards pumping it into them. The intensity was determined by measuring the spectral power of the speech signal. Figure 2 shows the spectral power and the pitch variation of normal and esophageal speech. It is clear that, esophageal speech allows larger variations of intensity. However, fundamental frequency is chaotic and difficult to detect. These unstable variations of intensity and F0 are responsible of the poor audio quality of ES.

For this reason, several approaches have been proposed to improve the quality and intelligibility of the alaryngeal speech. To enhance the quality of esophageal speech, Qi attempted replacing the voicing source of esophageal speech using a LPC method [7], [8], [9]. In [10] the authors proposed to use a simulated glottal waveform and a smoothed F0 to enhance tracheo-esophageal speech (TE). In order to reduce breath and hardness of the original speech, [11] used a synthetic glottic waveform and a model for jitter and reflection reduction. For synthesizing a laryngeal voice from the whispered voice, [12] proposed a

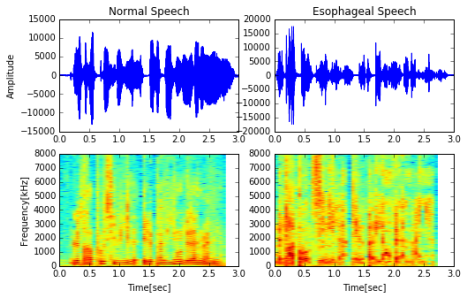

Figure 1: Example of waveforms and spectrograms of both normal and esophageal speech

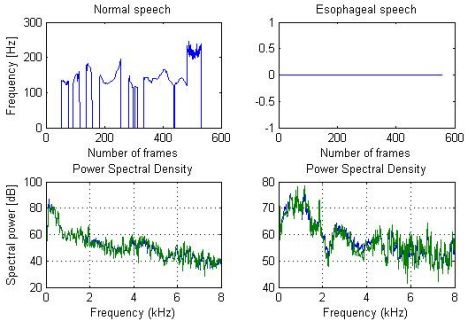

Figure 2: Example of F0 contours and spectral power of both normal and esophageal speech

Mixed-Excitation Linear Prediction (MELP) which consists in using formant structure modification and pitch estimation. This technique does not work in real time and furthermore the unvoiced phonemes are not converted. In order to produce more natural speech, [13] used a linear code excitation prediction (CELP) for estimating the pitch contours from whispered voice. Other approaches have been tempted to enhance pathological speech, based on transformations of their acoustic characteristics. Such as: combined root cepstral subtraction and spectral subtraction procedure for denoising electrolarynx speech [14]; the use of formant synthesis [3]; comb filtering [4]; the use of LPC for approximating the vocal tract filter [15] .

To increase the quality of TE speech, del Pozo and Young [16] proposed to estimate the new durations of TE phones with a prediction by regression trees constructed from laryngeal data.

To enhance alaryngeal speech, Tanaka et al. [17] proposed a spectral subtraction to reduce noise and a statistical voice conversion method to predict excitation parameters. Doi et al. [18] proposed to convert alaryngeal speech to be perceived as pronounced by a speaker with laryngeal voice.

However, all the conversion methods proposed are often quite complex and in addition they can generate errors in the parameters estimation. As a result these methods produce artificial synthetic speech because of the absence of realistic excitation signals estimation.

That is why, we propose a new algorithm for enhancing ES using a new voice conversion technique, based on estimating cepstral excitation (and phase) by a search of realistic examples in the target training space.

2.2. Voice conversion algorithm based on cepstral excitation prediction

We describe in the sequel a conversion method based on cepstral excitation prediction. In the proposed algoritm, vocal tract and excitation coefficients are separately estimated. The proposed method consists of a training and conversion phase. The training phase consists of three stages: speech analysis or feature extraction, alignment and computation of a mapping function. The conversion procedure consists of conversion of cepstral parameters, excitation and phase prediction and synthesis.

2.3. Feature extraction

In order to convert esophageal speech into normal speech, we use two parallel corpora, the first one from the source speaker (esophageal speech) and the second one from the target speaker (laryngeal speech). These corpora undergo a parameterization process, which consists in extracting cepstral feature vectors. To estimate spectral/cepstral features, different features have been considered by several researchers: Log spectrum was used in [19]; Mel-cepstrum (MCEP) was used in [20], [21], [35]; Mel-frequency cepstral coefficients MFCC were used in [22]. In this work, we use real Fourier cepstrum [23]. Figure 3 details the different steps involved in transforming the speech signal into its cepstral domain representation.

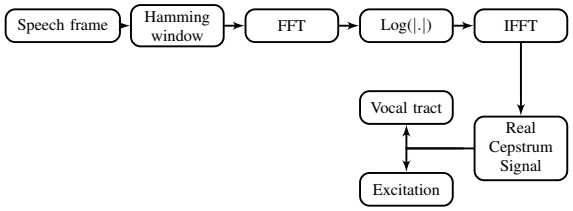

Figure 3: Bloc diagram of cepstrum based feature extraction

The linguistic contents of the speech signal are encoded into a set of coefficients:

- The first cepstral coefficient c_0
- The vocal tract cepstral vector $[c_1....c_{32}]$
- The cepstral excitation $[c_{33}....c_{256}]$
- The phase coefficients $[p_0....p_{256}]$

So we must find transformation prediction methods for the c_0, the vocal tract vector, the excitation vector and the phase coefficients.

2.4. Alignment

Dynamic time warping DTW [24], [33] is used to find an optimal alignment between the source sequences of vectors X=[X_1,X_2,...,X_s] and target sequences of vectors Y=[Y_1,Y_2,...,Y_t]. To align these two sequences of vectors, we must find an alignment path (A,B) where A=[a_1,a_2,...,a_U] and B=[b_1,b_2,...,b_U] are sequences of indices used to align X and Y. X and Y aligned are given by X=[X_{a_1},X_{a_2},...,X_{a_U}] and Y=[Y_{b_1},Y_{b_2},...,Y_{b_U}].

2.5. Training of the conversion function

To train the mapping function F, the aligned sequence of vocal tract vectors are used. The mapping function is estimated using a Gaussian Mixture Model (GMM) [25] (our baseline system), or a Deep Neural Network (DNN) [26].

2.5.1. Gaussian Mixture Model

The joint probability of vector z, which is the concatenation of a source vector x and its mapped target vector y, is represented as the sum of G multivariate Gaussian densities, given by:

$$p(z) = \sum_{i=1}^{G} \alpha_i N_i(z, \mu_i, \Sigma_i) \quad (1)$$

where $N_i(z, \mu_i, \Sigma_i)$ denotes the i-th Gaussian distribution with a mean vector μ_i and a covariance matrix Σ_i. G represents the total number of mixture components and α_i is the mixture weight:

$$\alpha_i = \frac{N_{s,i}}{N_s} \quad (2)$$

where $N_{s,i}$ and N_s are respectively the number of vectors in class i and the total number of source vectors classified. the mean vector μ_i is calculated as follow:

$$\mu_i^z = \frac{\sum_{k=1}^{N_{s,i}} z_i^k}{N_{s,i}} \quad (3)$$

The conversion function is then defined as a regression $E[y/x]$ given by formula 4:

$$F(x) = E[y/x] = \sum_{i=1}^{Q} p(i/x)(\mu_i^y + \Sigma_i^{yx}(\Sigma_i^{xx})^{-1}(x - \mu_i^x)) \quad (4)$$

where

$$\Sigma_i^z = \begin{bmatrix} \Sigma_i^{xx} & \Sigma_i^{xy} \\ \Sigma_i^{yx} & \Sigma_i^{yy} \end{bmatrix} and\ \mu_i^z = \begin{bmatrix} \mu_i^x \\ \mu_i^y \end{bmatrix}$$

Formula 4 represents the mapping function that transforms a source cepstum vector x into its corresponding target cepstrum vector $F(x)$.

2.5.2. Deep neural network

We propose to use a Multi Layer Perceptron (MLP) [20], [29]. A DNN is a feed-forward neural network with many hidden layers. Most studies have applied a single network on spectral envelopes [21], [23]. We further apply a deep neural network in order to transform a source voice into a target voice. The goal of a DNN is to approximate some function f(.) defined as:

$$y = f(x, \theta) \quad (5)$$

by learning the value of the parameters θ giving the best function approximation mapping all inputs x to outputs y. For each hidden unit i, an activation function g(.) is used to map the inputs from the layer x_i to an outputs y_i.

$$y_i = f(x_i) \quad (6)$$

where

$$x_i = b_i + \Sigma_k y_k w_{ki} \quad (7)$$

and b_i is the bias of unit; k is the unit index of layer; w_{ki} is the weight of the connetion.

Recently Rectified Linear Unit ReLU activation function has become more and more popular [30], [31], [32] for its simplicity and good performance. In this work, we choose a deep feedforward network with ReLU activation function. To reduce the problem of DNN over-fitting we use the Dropout technique [36]. Dropout allows to give up units (hidden and visible) in a deep neural network. Consequently, it prevents deep neural network from over-fitting by providing a way of approximately combining exponentially many different deep neural network architectures efficiently.

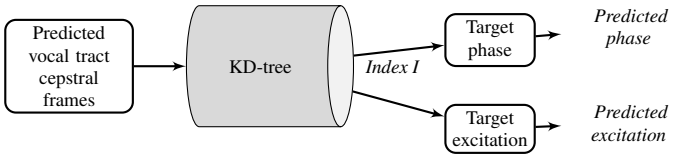

Figure 4: KD-Tree query for excitation and phase prediction

2.6. Conversion

At this stage, cepstral coefficients are extracted from the esophageal speech signal. Then only the vocal tract vectors $X_i = [X_i^1 ... X_i^{32}]$ are converted by the previously described methods.

2.7. Synthesis

The main contribution of our approach consists in the prediction of the cepstral excitation and phase using a KD-Tree [27], [34]. As shown in Figure 4, converted vocal tract cepstral vectors are used to estimate cepstral excitation and phase coefficients.

The binary KD-Tree is constructed with concatenated target vocal tract cepstral vectors of length NB_{frame}. The KD-Tree is queried by the predicted vocal tract cepstral vectors previously concatenated as a frame of length NB_{frame}. This query provides an index I which corresponds to the nearest target vocal tract cepstral frame.

The target cepstral excitation vectors are concatenated to form a frame of length NB_{frame}. In the same manner the target phase vectors are concatenated.

Thereafter, index I is used as the index of a desired cepstral excitation and phase frame. Since excitation and phase are estimated, a complex spectrum is formed using magnitude and phase spectra. The spectral synthesizer we use is based on overlapping and adding the inverse Fourier transform temporal signals using OLA-FFT. To preserve the characteristics of the vocal tract of the source speaker the first cepstral packet (vocal tract packet), in one of our experiments, has not been modified at the resynthesis stage.

3. Experimental evaluations

3.1. Datasets

The Datasets have been created by 3 French speakers: two laryngectomees (PC and MH), and a speaker AL with normal voice. For each speaker 289 phonetically balanced utterances have been recorded. The speech of each corpus is sampled at 16 kHz.

3.2. Experimental conditions

All conditions of the experiments are summarized in Table 1 :

3.3. Objective evaluations

To evaluate our system, we have chosen to calculate the Log Spectral Distortion and Signal to Error Ratio (SER) . The Log Spectral Distortion (LSD), via cepstrum representation becomes the cepstral distance CD [28].

$$CD(x, y)\,[dB] = \frac{10}{log10}\sqrt{\sum_k (c_k(x) - c_k(y))^2} \quad (8)$$

Table 1: *EXPERIMENTAL CONDITIONS.*

Number of GMMs	64
Window length	32 ms
Shift length	4 ms
Number of training utterances	200
Number of test utterances	20
FFT size	512
NB_{frame}	20
DNN structure	512*5
Number of epochs	100

where $c_k(x)$ and $c_k(y)$ are respectively the k-th cepstral coefficient of converted and target cepstrum vectors.

SER is represented by the following formula

$$SER\,[dB] = -10 * log_{10} \frac{\sum_k ||y_k - \hat{y_k}||^2}{\sum_k ||y_k||^2} \qquad (9)$$

where y_k and $\hat{y_k}$ are respectively the target and converted cepstral vectors. Tables 2 and Table 3 show the different values of Signal to Error Ratio (SER) and cepstral Distance (CD) between the source and target cepstrum, then between predicted and target cepstrum.

Table 2: *SER [dB]*

Methods	PC	MH
Extracted	2.7	2.97
GMM-based method	12.33	11.39
DNN-based approach	12.99	11.80

Table 3: *CD [dB]*

Methods	PC	MH
Extracted	9.28	9.03
GMM-based method	5.37	5.53
DNN-based approach	5.29	5.28

We can observe that the cepstral vectors of esophageal speech are very different from those of normal speech. We can see also that the proposed conversion methods can take into accound these large differences. More significantly, it is clear that the proposed DNN-based approach performs much better than the GMM-based method for the voice conversion task.

3.4. Perceptual evaluations

We conducted two opinion tests: one for naturalness and the other one for intelligibility. The following four types of speech samples were evaluated by ten listeners.

- ES (esophageal speech)
- GMM_CVT: The vocal tract cepstral vectors are converted using the GMM model, then the converted vocal tract cepstral vectors are used to estimate excitation and phase. The converted vocal tract vectors are used in the synthesis process.
- DNN_CVT: The vocal tract cepstral vectors are converted using the DNN model, and the converted vocal tract cepstral vectors are used in the synthesis process.

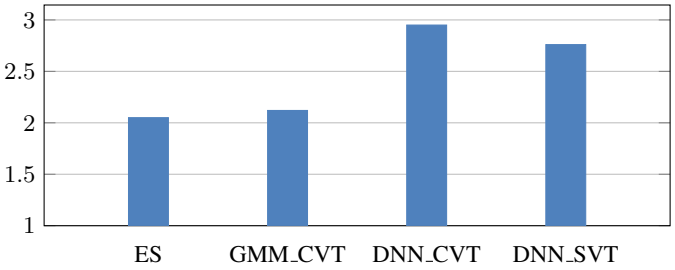

Figure 5: Mean opinion scores on naturalness

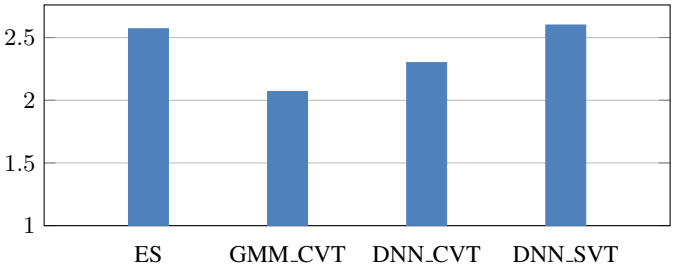

Figure 6: Mean opinion scores on intelligibility

- DNN_SVT: We use the source vocal tract cepstral vectors in the synthesis process.

Each listener evaluated 32 samples in each of the two tests. Figures 5 and 6 show the results for naturalness and intelligibility tests. The rating scale was a standard MOS (Mean Opinion Score) scale (1=bad 2=poor 3=fair 4=good 5=excellent). The results show that DNN methods perform better than GMM method and these methods are expected to have a fairly good acceptance among laryngectomees. Finally, about 70 % of the listeners participating in the subjective evaluation preferred the use of the proposed system based on preserving the source vocal tract cepstral vectors. The other listeners preferred DNN_CVT: the results provided by this method seem to them more natural than those provided by DNN_SVT. Some samples obtained by this work are presented in the following web link: demo.
These evaluation results show that our proposed system is very effective for improving the naturalness of esophageal speech while preserving its intelligibility.

4. Conclusions and prospects

In this article, we propose a voice conversion algorithm for esophageal speech enhancement. The originality of our approach lies in the prediction of cepstral excitation and phase using a KD-Tree. The vocal tract cepstral vectors are converted using two methods, one based on DNN and the other one on GMM, and these converted vectors are used for predicting excitation and phase. But in the synthesis stage (in experiment DNN_SVT) those converted vectors are not used in order to retain the vocal tract characteristics of the source speaker. Objective and subjective evaluations have demonstrated that our method provides significant improvements in the quality of the converted esophageal speech while preserving its intelligibility.

In a near future we intend to further increase the naturalness of the transformed esophageal speech and consequently its intelligibility.

5. References

[1] H. Doi, K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano, "Statistical approach to enhancing esophageal speech based on Gaussian mixture models", Acoustics Speech and Signal Processing(ICASSP), pp. 4250-4253, 2010.

[2] H. Doi, K. Nakamura, T. Toda, H. Saruwatari, K. Shikano, "Enhancement of Esophageal Speech Using Statistical Voice Conversion", in APSIPA 2009, Sapporo, Japan, pp. 805-808, Oct. 2009.

[3] K. Matsui, N. Hara, N. Kobayashi, H. Hirose, "Enhancement of esophageal speech using formant synthesis", Proc. ICASSP, pp. 1831-1834, 1999-May.

[4] A. Hisada and H. Sawada. "Real-time clarification of esophageal speech using a comb filter", Proc. ICDVRAT. 2002.

[5] T. Toda, A. Black and K. Tokuda, "Voice Conversion Based on Maximum-Likelihood Estimation of Spectral Parameter Trajectory", IEEE Transactions on Audio, Speech and Language Processing, vol. 15, no. 8, pp. 2222-2235, 2007.

[6] Y. Stylianou, O. Cappé and E. Moulines, "Continuous probabilistic transform for voice conversion", IEEE Transactions on Speech and Audio Processing, vol. 6, no. 2, pp. 131-142, 1998.

[7] Ning Bi and Yingyong Qi, "Application of speech conversion to alaryngeal speech enhancement", IEEE Transactions on Speech and Audio Processing, vol. 5, no. 2, pp. 97-105, 1997.

[8] Y. Qi and B. Weinberg, "Characteristics of Voicing Source Waveforms Produced by Esophageal and Tracheoesophageal Speakers", Journal of Speech Language and Hearing Research, vol. 38, no. 3, p. 536, 1995.

[9] Y. Qi, "Replacing tracheoesophageal voicing sources using LPC synthesis", The Journal of the Acoustical Society of America 88.3,pp 1228-1235,1990.

[10] Y. Qi, B. Weinberg and N. Bi, "Enhancement of female esophageal and tracheoesophageal speech", The Journal of the Acoustical Society of America, vol. 98, no. 5, pp. 2461-2465, 1995.

[11] A. del Pozo and S. Young,"Continuous Tracheoesophageal Speech Repair", In Proc. EUSIPCO, 2006.

[12] H. I. Trkmen and M. E. Karsligil. "Reconstruction of dysphonic speech by melp", Iberoamerican Congress on Pattern Recognition. Springer, Berlin, Heidelberg, 2008.

[13] H. Sharifzadeh, I. McLoughlin and F. Ahmadi, "Reconstruction of Normal Sounding Speech for Laryngectomy Patients Through a Modified CELP Codec", IEEE Transactions on Biomedical Engineering, vol. 57, no. 10, pp. 2448-2458, 2010.

[14] D. Cole, S. Sridharan, M. Moody, S. Geva, "Application of Noise Reduction Techniques for Alaryngeal Speech Enhancement", Proc. of The IEEE TENCON, pp. 491-494, 1997.

[15] B. García, J. Vicente, and E. Aramendi. "Time-spectral technique for esophageal speech regeneration", 11th EUSIPCO (European Signal Processing Conference). IEEE, Toulouse, France. 2002.

[16] A. del Pozo and S. Young. "Repairing Tracheoesophageal Speech Duration", In Proc. Speech Prosody, 2008.

[17] K. Tanaka, T. Toda, G. Neubig, S. Sakti, and S. Nakamura, "A hybrid approach to electrolaryngeal speech enhancement based on noise reduction and statistical excitation generation", IEICE Trans. on Inf. and Syst., vol. E97-D, no. 6, pp. 14291437, Jun. 2014.

[18] H. Doi and al. "Alaryngeal speech enhancement based on one-to-many eigenvoice conversion", IEEE/ACM Transactions on Audio, Speech, and Language Processing 22.1, pp 172-183,2014.

[19] F.-L. Xie, Y. Qian, Y. Fan, F. K. Soong, and H. Li, "Sequence error (SE) minimization training of neural network for voice conversion", in Proc. Interspeech, 2014.

[20] D. Srinivas, E. V. Raghavendra, B. Yegnanarayana, A. W. Black, K. Prahallad, "Voice conversion using artificial neural networks", Proc. IEEE Int. Conf. Acoust. Speech Signal Process., pp. 3893-3896, Apr. 2009.

[21] Ling-Hui Chen, Zhen-Hua Ling, Li-Juan Liu and Li-Rong Dai, "Voice Conversion Using Deep Neural Networks With Layer-Wise Generative Training", IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 22, no. 12, pp. 1859-1872, 2014.

[22] T. Nakashika, T. Takiguchi and Y. Ariki, "Voice Conversion Using RNN Pre-Trained by Recurrent Temporal Restricted Boltzmann Machines", IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 23, no. 3, pp. 580-587, 2015.

[23] T. Nakashika, R. Takashima, T. Takiguchi, Y. Ariki, "Voice conversion in high-order eigen space using deep belief nets", pp. 369-372, 2013.

[24] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition", IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 26, no. 1, pp. 43-49, 1978.

[25] H. Valbret, "Système de conversion de voix pour la synthèse de parole", Diss. 1993.

[26] Z.-H. Ling, S.-Y. Kang, H. Zen, A. Senior, M. Schuster, X.-J. Qian, H. Meng, L. Deng, "Deep learning for acoustic modeling in parametric speech generation: A systematic review of existing techniques and future trends", IEEE Signal Process. Mag., vol. 32, no. 3, pp. 35-52, May 2015.

[27] S. Arya, "Nearest neighbor searching and applications", Ph.D. thesis, Univ. of Maryland at College Park, 1995.

[28] M. M. Deza, E. Deza, "Encyclopedia of Distances", Berlin, Springer, 2009.

[29] S. Desai, A. Black, B. Yegnanarayana, K. Prahallad, "Spectral mapping using artificial neural networks for voice conversion", IEEE Trans. Audio Speech Lang. Process., vol. 18, no. 5, pp. 954-964, Jul. 2010.

[30] V. Nair and G.E. Hinton, "Rectified linear units improve restricted boltzmann machines", Proceedings of the 27th international conference on machine learning (ICML-10). 2010.

[31] A. L. Maas, A. Y. Hannun, A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models", Proc. ICML, vol. 30, 2013.

[32] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification", arXiv preprint arXiv:1502.01852, 2015.

[33] M. Muller, "Information retrieval for music and motion", Vol. 2. Heidelberg, Springer, 2007.

[34] K. Zhou, Q. Hou, R. Wang and B. Guo, "Real-time KD-tree construction on graphics hardware", ACM Transactions on Graphics, vol. 27, no. 5, p. 1, 2008.

[35] L. Sun, S. Kang, K. Li, and H. Meng, "Voice conversion using deep bidirectional long short-term memory based recurrent neural networks", Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on. IEEE, 2015.

[36] Nitish Srivastava et al. "Dropout: a simple way to prevent neural networks from overfitting." Journal of machine learning research 15.1 (2014): 1929-1958.